View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Wageningen University & Research Publications

D1250–D1254 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 1 December 2011*
doi:10.1093/nar/gkr1099

MetaBase—the wiki-database of biological databases

Dan M. Bolser[1,*], Pierre-Yves Chibon[2], Nicolas Palopoli[3], Sungsam Gong[3], Daniel Jacob[4], Victoria Dominguez Del Angel[5], Dan Swan[6], Sebastian Bassi[7], Virginia González[3], Prashanth Suravajhala[8,*], Seungwoo Hwang[9], Paolo Romano[10], Rob Edwards[11], Bryan Bishop[1,*], John Eargle[12], Timur Shtatland[13], Nicholas J. Provart[14], Dave Clements[15], Daniel P. Renfro[16], Daeui Bhak[17] and Jong Bhak[1,18,*]

[1]Personal Genomics Institute, Genome Research Foundation, Suwon, 443-270, South Korea, [2]Plant Breeding, Wageningen University, Wageningen, The Netherlands, [3]Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Bernal, Buenos Aires, Argentina, [4]INRA, UMR 1332, Fruit Biology and Pathology Centre. Bordeaux, BP 81, F-33140 Villenave d'Ornon, [5]Institut National de la Recherche Agronomique, URGI, Route de Saint Cyr 78026, Versailles, France, [6]Oxford Gene Technology, Begbroke Science Park, Sandy Lane, Yarnton, Oxford, OX5 1PF, UK, [7]Genes Digitales, Buenos Aires, Argentina, [8]Bioinformatics Organization, 225 Cedar Hill Street, Suite 200 Marlborough, MA 01752, USA, [9]Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon, Korea, [10]IRCCS AOU San Martino-IST National Cancer Research Institute, Largo R. Benzi 10, I-16132, Genova, Italy, [11]Department of Biology and Department of Computer Sciences, San Diego State University, San Diego, CA 92182, [12]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, [13]http://ksvetu.blogspot.com/, Melrose, USA, [14]Department of Cell & Systems Biology, Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON, Canada, [15]Department of Biology, Emory University, Atlanta, Georgia, [16]Department of Biochemistry and Biophysics, Texas A&M University and Texas Agrilife Research, USA, [17]Interdisciplinary Research Program of Bioinformatics and Longevity Science, Pusan National University, Busan, Korea, [18]Theragen BiO Institute, Suwon 443-270, South Korea

Received August 22, 2011; Revised and Accepted November 3, 2011

ABSTRACT

Biology is generating more data than ever. As a result, there is an ever increasing number of publicly available databases that analyse, integrate and summarize the available data, providing an invaluable resource for the biological community. As this trend continues, there is a pressing need to organize, catalogue and rate these resources, so that the information they contain can be most effectively exploited. MetaBase (MB) (http://MetaDatabase.Org) is a community-curated database containing more than 2000 commonly used biological databases. Each entry is structured using templates and can carry various user comments and annotations. Entries can be searched, listed, browsed or queried. The database was created using the same MediaWiki technology that powers Wikipedia, allowing users to contribute on many different levels. The initial release of MB was derived from the content of the 2007 *Nucleic Acids Research* (*NAR*) Database Issue. Since then, approximately 100 databases have been manually collected from the literature, and users have added information for over 240 databases. MB is synchronized annually with the static Molecular Biology Database Collection provided by *NAR*. To date, there have been 19 significant contributors to the project; each one is listed as an author here to highlight the community aspect of the project.

*To whom correspondence should be addressed. Tel: + 01223 968 518; Email: dan.bolser@gmail.com
Correspondence may also be addressed to Bryan Bishop. Tel: 1-512-203-0507; Email: kanzure@gmail.com
Correspondence may also be addressed to Jong Bhak. Tel: 031-888-9311; Fax: 031-888-9314; Email: jongbhak@yahoo.com

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://nar.oxfordjournals.org/ at Wageningen UR Library on July 26, 2013

INTRODUCTION

When discussing biological databases, there are simply too many different resources to comprehensively cover the topic in a short introduction. There are well-established data warehouses that act as community repositories for data of a single type such as GenBank (1), PDB (2) and ArrayExpress (3). There are organism-specific databases, combining many different types of data under a unifying, genomic framework such as TAIR (4), FlyBase (5) and WormBase (6). There are databases of derived data, collecting and systematizing the body of knowledge from the scientific literature such as GTEx (http://www.ncbi.nlm.nih.gov/gtex/GTEX2/gtex.cgi), TRANSFAC (7), Brenda (8) and ChEMBL (9). There are competing databases that cover specific kinds of -omics information, collecting data from different experiments within a common biological theme such as DIP (10), HPID (11) and IntAct (12). There are classification databases (13,14), databases of terminology (15,16), databases of protein families (17,18) and databases built around diseases (19) or taxonomic groups (20). This list barely scratches the surface, but gives a flavour of the number, types and diversity of biological databases.

As the type and volume of biological data continues to increase, so do the type and number of databases that analyse, integrate and summarize the available data. For example, querying the database of biomedical publications PubMed (21) shows that the number of unique publications with the word 'database' in the title has increased from just 2 in 1980 to 91 in 1990 and 469 in 2000. Since 1990, there has been an exponential increase in the number of database publications per year, reaching over 1000 per year between 2008 and 2010 (Figure 1). If this trend continues, the number of database publications per year will double to nearly 2000 by 2015.

Biological databases have proven crucially important for basic research, however, the current growth in the available databases creates several problems. Researchers seeking the most up-to-date and comprehensive information in their domain may struggle to identify the definitive sources of reliable data from among the many resources available. Initially, it is difficult to judge the strengths, weaknesses, or status of the available resources without peer guidance. For these reasons, the proliferation of resources may, ironically, lead to an increase in redundancy, as new resources are created to cope with the perceived problems or omissions of existing databases. This process is exacerbated by a lack of public forums where researchers can engage database creators to discuss databases and suggest improvements.

These issues have created an unfortunate situation whereby many resources are short-lived, existing for only a short time before being abandoned. This 'half-life' is analogous to 'link rot' (22). This creates a vicious cycle, whereby the publication of database resources is devalued (23). To address these problems, we have created MetaBase (MB), a wiki-based database of biological databases.

DATABASE DESCRIPTION

MB is a community-curated database of all the biological databases available on the Internet. The aim of the project is to make it easy for researchers to quickly find relevant information about useful databases. Entries can be searched, queried or browsed by category, and users can contribute, update and maintain the data in many different ways. Each database in MB is described in a semi-structured way using forms and templates. Entries carry data for various fields and allow a free-text description of the resource. In detail, data for each database include a brief description, a URL, a contact email, links to associated literature and various categorization tags. In addition, entries can carry various user comments and annotations.

MB has been implemented using MediaWiki (MW), the same software that powers Wikipedia, probably the best known user-contributed resource in the world (http://wikipedia.org). The MediaWiki system allows users to contribute to the project on many different levels, ranging from authors and editors to curators and site designers. Within the MW system, we created one wiki-page per database entry. The information about each database is structured by using a template with named fields. The template stores data for each database

Downloaded from http://nar.oxfordjournals.org/ at Wageningen UR Library on July 26, 2013

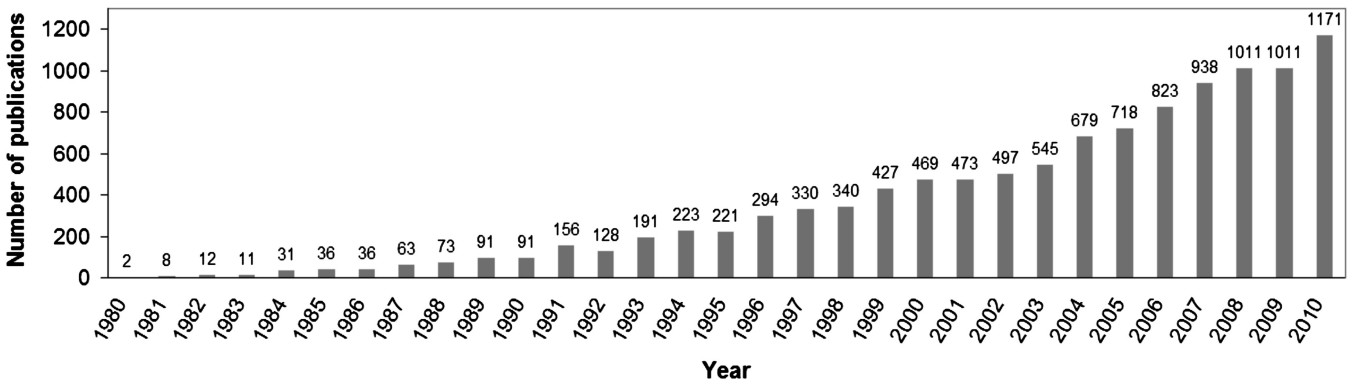

Figure 1. The growth in the number of database publications per year. Each bar shows the number of research articles with the keyword 'database' appearing in the article title in the given year. The count only covers articles indexed in PubMed. The increase shows an exponential trend that will produce nearly 2000 database publications per year by 2015.

internally using the Semantic MediaWiki extension (http://semantic-mediawiki.org), allowing data to be queried within the wiki directly, by additional extensions or via the semantic web. In particular, we use the Semantic Forms extension (http://www.mediawiki.org/wiki/SF) to allow users to create or edit entries and the Semantic Drilldown extension (http://www.mediawiki.org/wiki/SD) to allow users to explore the database. User comments are collected as free text, just like in Wikipedia.

FEATURES

The MW platform provides a robust base from which to build an online resource. By using MW, many powerful features are provided 'for free'. The use of MW to support Wikipedia demonstrates the scalability and security of the system, guaranteeing developer support and providing a degree of familiarity to users. Out of the box, MW provides searching, editing, versioning, history and discussion features, as well as user account management and user-email functions. MW includes a powerful extension framework for easily adding functionality.

One criticism of MW is that it provides largely unstructured information, not suitable for advanced searching or reporting. To this end, we employ Semantic MediaWiki and Semantic Forms to create a wiki-database system suitable for maintaining a user-contributed database of information.

DATABASE CONTENTS

Currently, there are 1795 entries in MB, each describing a different biological database. The initial release was derived from the content of the 2007 Nucleic Acids Research (*NAR*) Database Issue (24). Specifically, each database page was 'seeded' with text from the Molecular Biology Database Collection provided by *NAR* (25). Subsequent releases have been updated into MB on a semi-regular basis. Since the initial release, there have been over 100 user contributed resources added, in addition to 100 resources that were manually collected from the literature. Most of these were taken from database publications in BMC Bioinformatics and BMC Biology. To date, there have been 19 significant contributors to the project, each of whom has been listed as an author on this publication. This step was taken to highlight the community aspect of the MB project. The homepage has been visited approximately 100 000 times. The project has 80 registered users in total, and there have been approximately 15 000 edits. We hope that with ongoing improvements and through increased publicity, usage will continue to grow helping to establish MB as a powerful and referential community resource.

FUTURE DIRECTIONS

In the future, we hope to use MB as a resource to allow more communication between database developers and user communities, acting as a common portal for the biological database community. To achieve this goal, we will automatically register the database's contact email address and add the database's discussion page to that user's 'watch list'. Comments will then automatically alert the contact, providing them with the opportunity to reply. We hope to add user rating functionality and usage statistics to each resource. This will be done with a combination of existing MediaWiki extensions, adding links to social networking sites and automatic queries to collect the number of citations for each resource. We expect that MB could be used as a source of genuine metadata for data integration projects, and we plan to incorporate ontologies such as EDaM (26,27) and the Biomedical Resource Ontology (28), and to develop links with similar projects such as BioCatalogue (29) and BioDBCore (30).

Finally, we aim to improve the content of MB through an aggressive marketing strategy, contacting the relevant mailing lists, forums and news groups, as well as exploiting the collection of contact email addresses, thereby encouraging the community to contribute to the maintenance of this important resource.

RELATED WORK

MB is by no means unique. There are many related resources, falling into two broad categories: 'BioWikis' and 'databases of biological databases'.

First, there are several other 'BioWiki' projects. Like MB, these projects use the tremendously successful MediaWiki software platform to provide user-contributed content to the biological community. For a comprehensive list of important and interesting BioWiki projects, see the BioWiki database on Bioinformatcs.Org (http://bioinformatics.org/wiki/BioWiki). The most successful collection of user-contributed content is Wikipedia (http://www.wikipedia.org/). The success of Wikipedia is intimately related to the success of the MediaWiki software platform, leading to a proliferation of wikis, including several BioWiki projects. However, Wikipedia is still a very important resource for biologists (e.g. http://en.wikipedia.org/wiki/Wikipedia:MCB). Wikipedia maintains a sizeable list of biological databases (http://en.wikipedia.org/wiki/List_of_biological_databases), and many of the databases in MB also have articles in Wikipedia.

Second, there are several 'databases of biological databases', which aim to provide a list of all the most important biological databases and data resources available on the Internet. Several prominent biological database collections and related projects are listed in Table 1 (see also http://metadatabase.org/wiki/Help:Related).

DISCUSSION

Biological databases have proven crucially important for basic research. However, exponential growth in the volume of biological data has led to several problems. MB is an international, community-based database that aims to list all the commonly used biological databases in the world. Here, we have created a new scientific-wiki that addresses some of the issues described earlier. The first

Downloaded from http://nar.oxfordjournals.org/ at Wageningen UR Library on July 26, 2013

Table 1. Projects with a similar scope to MB

Name	Description	URL
The Molecular Biology Database Collection	A public on-line resource that lists the databases described in Nucleic Acids Research, together with other databases of value to the biologist (25).	http://www.oxfordjournals.org/nar/database/c/
OBRC: Online Bioinformatics Resources Collection	Contains annotations and links for 1746 bioinformatics databases and software tools.	http://www.hsls.pitt.edu/guides/genetics/obrc/
The Bioinformatics Links Directory	Features curated links to molecular resources, tools and databases (31).	http://bioinformatics.ca/links_directory/
CABRI: Common Access to Biotechnology Resources and Information	An service to search European Biological Resource Centre catalogues. The catalogues may be searched independently, or as one, and the located materials ordered online or by post (32).	http://www.cabri.org/
DBD: Database of Biological Database	Consists of 1200 database entries covering wide range of databases useful for biological researchers.	http://www.biodbs.info/
BioDBCore	A community-defined description of the core attributes of biological databases (28).	http://biocurator.org/biodbcore.shtml
MetaBasis	A database of metadata for bioinformatics software tools and databases. The system contains 3229 published bioinformatics tools and databases (33).	http://bioserver-1.bioacademy.gr/Metabasis/
Biomed Central Databases	A catalogue of online databases with more than 1100 sites covering a wide range of biomedical topics.	http://databases.biomedcentral.com/
OReFiL	An Online Resource Finder for Life sciences (34).	http://orefil.dbcls.jp/
NIST Data Gateway	Provides easy access to many of the The National Institute of Standards and Technology databases, covering a many different scientific disciplines.	http://srdata.nist.gov/gateway/

These projects aim to list the most important biological databases and data resources available on the Internet. For a version of this table that you can edit, see http://metadatabase.org/wiki/Help:Related

version of the system was based on a static database of biological databases that has been imported to a wiki system for community annotation. Although similar to several other 'lists of resources', MB is unique, being the only truly user-editable list of databases. The *NAR* Molecular Biology Database Collection is a curated database with strict criteria for inclusion. It covers only a relatively small number of the available molecular biology databases (M. Galperin, personal communication). In contrast, we hope MB, with its liberal wiki-based inclusion policy, might be useful as a wider, more general list with quicker updates.

ACKNOWLEDGEMENTS

MB was first hosted on the Bioinformation Objects Community Cluster (BiO.CC) and was created as a competition entry hosted by BiO.CC. Currently, MB is hosted at Bioinformatics.Org. D.M.B. would like to thank J.B. and Jeff Bizzaro for hosting and all the contributors to MB.

FUNDING

Industrial Strategic technology development program, (10040231), ''Bioinformatics platform development for next generation bioinformation analysis'' funded by the Ministry of Knowledge Economy (MKE, Korea). Funding for Open access charge: Genome Research Foundation's internal Biowiki funds.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2011) GenBank. *Nucleic Acids Res.*, **39**, D32–D37.
2. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
3. Parkinson,H., Sarkans,U., Kolesnikov,N., Abeygunawardena,N., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Holloway,E. *et al.* (2011) ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **39**, D1002–D1004.
4. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
5. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al.*; FlyBase Consortium (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.*, **37**, D555–D559.
6. Rogers,A., Antoshechkin,I., Bieri,T., Blasiar,D., Bastiani,C., Canaran,P., Chan,J., Chen,W.J., Davis,P., Fernandes,J. *et al.* (2008) WormBase 2007. *Nucleic Acids Res.*, **36**, D612–D617.
7. Knüppel,R., Dietze,P., Lehnberg,W., Frech,K. and Wingender,E. (1994) TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. *J. Comput. Biol.*, **1**, 191–198.
8. Scheer,M., Grote,A., Chang,A., Schomburg,I., Munaretto,C., Rother,M., Söhngen,C., Stelzer,M., Thiele,J. and Schomburg,D. (2011) BRENDA, the enzyme information system in 2011. *Nucleic Acids Res.*, **39**, D670–D676.
9. Gaulton,A., Bellis,L.J., Bento,A.P., Chambers,J., Davies,M., Hersey,A., Light,Y., McGlinchey,S., Michalovich,D., Al-Lazikani,B. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.

Downloaded from http://nar.oxfordjournals.org/ at Wageningen UR Library on July 26, 2013

10. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
11. Han,K., Park,B., Kim,H., Hong,J. and Park,J. (2004) HPID: the Human Protein Interaction Database. *Bioinformatics*, **20**, 2466–2470.
12. Aranda,B., Achuthan,P., Alam-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.
13. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
14. Cuff,A.L., Sillitoe,I., Lewis,T., Clegg,A.B., Rentzsch,R., Furnham,N., Pellegrini-Calace,M., Jones,D., Thornton,J. and Orengo,C.A. (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res.*, **39**, D420–D426.
15. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
16. Gene Ontology Consortium. (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**, D331–D335.
17. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
18. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
19. Forbes,S.A., Bindal,N., Bamford,S., Cole,C., Kok,C.Y., Beare,D., Jia,M., Shepherd,R., Leung,K., Menzies,A. *et al.* (2011) COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **39**, D945–D950.
20. Bombarely,A., Menda,N., Tecle,I.Y., Buels,R.M., Strickler,S., Fischer-York,T., Pujar,A., Leto,J., Gosselin,J. and Mueller,L.A. (2011) The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. *Nucleic Acids Res.*, **39**, D1149–D1155.
21. McEntyre,J. and Lipman,D. (2001) PubMed: bridging the information gap. *CMAJ*, **164**, 1317–1319.
22. Koehler,W. (2004) A longitudinal study of Web pages continued: a report after six years. *Informat. Res.*, **9**, paper 174.
23. Tan,T.W., Tong,J.C., Khan,A.K., de Silva,M., Lim,K.S. and Ranganathan,S. (2010) Advancing standards for bioinformatics activities: persistence, reproducibility, disambiguation and Minimum Information About a Bioinformatics investigation (MIABi). *BMC Genomics*, **11**, S27.
24. Bateman,A. (2007) Editorial. *Nucleic Acids Res.*, **35**, D1–D2.
25. Galperin,M.Y. and Cochrane,G.R. (2011) The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res.*, **39**, D1–D6.
26. Pettifer,S., Ison,J., Kalas,M., Thorne,D., McDermott,P., Jonassen,I., Liaquat,A., Fernández,J.M., Rodriguez,J.M. *et al.*; INB-Partners (2010) The EMBRACE web service collection. *Nucleic Acids Res.*, **38**, W683–W688.
27. Kalas,M., Puntervoll,P., Joseph,A., Bartaseviciute,E., Töpfer,A., Venkataraman,P., Pettifer,S., Bryne,J.C., Ison,J., Blanchet,C. *et al.* (2010) BioXSD: the common data-exchange format for everyday bioinformatics web services. *Bioinformatics*, **26**, i540–i546.
28. Tenenbaum,J.D., Whetzel,P.L., Anderson,K., Borromeo,C.D., Dinov,I.D., Gabriel,D., Kirschner,B., Mirel,B., Morris,T., Noy,N. *et al.* (2011) The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research. *J. Biomed. Inform.*, **44**, 137–145.
29. Bhagat,J., Tanoh,F., Nzuobontane,E., Laurent,T., Orlowski,J., Roos,M., Wolstencroft,K., Aleksejevs,S., Stevens,R., Pettifer,S. *et al.* (2010) BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Res.*, **38**, W689–W694.
30. Gaudet,P., Bairoch,A., Field,D., Sansone,S.-A., Taylor,C., Attwood,T.K., Bateman,A., Blake,J.A., Bult,C.J., Cherry,J.M. *et al.* (2011) Towards BioDBcore: a community-defined information specification for biological databases. *Nucleic Acids Res.*, **39**, D7–D10.
31. Chen,Y.B., Chattopadhyay,A., Bergen,P., Gadd,C. and Tannery,N. (2007) The Online Bioinformatics Resources Collection at the University of Pittsburgh Health Sciences Library System–a one-stop gateway to online bioinformatics databases and software tools. *Nucleic Acids Res.*, **35**, D780–D785.
32. Romano,P., Aresu,O., Manniello,M. and Parodi,B. (2004) Interoperability of CABRI Services and Biochemical Pathways Databases. *Comp. Funct. Genomics*, **5**, 169–172.
33. Atlamazoglou,V., Thireou,T., Hamodrakas,Y. and Spyrou,G. (2006) MetaBasis: a web-based database containing metadata on software tools and databases in the field of bioinformatics. *Appl. Bioinformatics*, **5**, 187–192.
34. Yamamoto,Y. and Takagi,T. (2007) OReFiL: an online resource finder for life sciences. *BMC Bioinformatics*, **8**, 287.

Downloaded from http://nar.oxfordjournals.org/ at Wageningen UR Library on July 26, 2013